文章编号:1001-8166(2011)04-0449-11

Web 时空数据挖掘研究进展*

孙 嘉[1,2],裴 韬[3],龚 玺[2,3],周成虎[1,3]

(1. 中国科学院烟台海岸带研究所,山东 烟台 264003; 2. 中国科学院研究生院,北京 100049;
3. 中国科学院地理科学与资源研究所,资源与环境信息系统国家重点实验室,北京 100101)

摘 要:随着互联网的迅速发展,Web 已经渗透到人类社会的各个角落,其中蕴含着大量关系社会、经济和生活的信息。从中挖掘出刻画事件时空范围的时空信息,可以为探索社会、自然事件以及行为主体的时空运动规律和知识提供丰富的素材。系统综述了 Web 时空数据挖掘的理论、方法和应用,首先介绍了 Web 时空数据挖掘的概念及分类,详细阐述了 Web 时空信息的特点和提取方法,其次针对 3 类 Web 时空数据挖掘的内容、方法及应用进行了综述,最后探讨了 Web 时空数据挖掘面临的难题、研究热点和未来领域的发展方向。

关 键 词:Web 时空数据挖掘;空间数据挖掘;时空信息;地理信息提取

中图分类号:TP311 **文献标志码**:A

1 引 言

互联网的迅猛发展使网络中累积的数据高速增加,而在这些网络信息中,刻画事件时空范围的时空信息在网络中无所不在,如网络新闻中的地名、网络广告中的商户地址等。Goodchild[1]认为,互联网是地理空间信息最大的收藏地,并且其中大部分都还未被利用。其原因有 3 个:①互联网作为社会信息传播的重要媒介之一,各个领域的实时新闻在网络中高速更新。其中,刻画新闻的要素——时空信息,在这些描述事件的文本或多媒体文件中无所不在。如地震发生的时间地点、重要会议起止的时间和开会地点等。②为了快速从互联网的海量信息池中获取信息,搜索引擎成为了连接用户和互联网的首要通道。而搜索引擎的使用日志则记录了不同用户的搜索条目和 IP 地址,是人类网络活动的真实反映。③近年来,以电脑、手机为客户端的空间定位信息服务(Location-based Servise ,LBS)有了快速发展。这些移动服务以地理位置为基础,将人类日常的空间活动反映在网络中。从上述这 3 个角度看,互联网将现实现象通过人类网络活动映射到虚拟世界中。Google Zeitgeist(http://www.google.com/zeitgeist)每年会发布全球热门搜索关键字排行榜,全球上升最快搜索关键词能够反映该年网民最关心的话题。在这种映射下,通过对互联网中海量的信息进行知识挖掘,获取的知识可以用来分析、指导、预测现实行为。

随着数据挖掘和知识发现(Data Mining and knowledge discovery, DM)的理论与技术发展,考虑空间邻近关系的空间数据挖掘(Spatial Data Mining, SDM)受到了广泛关注[2,3]。Han 等[4]在 1997 年提出了空间数据挖掘的原型系统 GEOMiner,李德仁等[5]在对 SDM 数据源、地理信息系统(GIS)数据挖掘和遥感图像数据挖掘的研究基础上,提出了云模型、数据场、地学粗空间和空间数据挖掘视角等新技术,构建了空间数据挖掘金字塔,并且导出空间观测

* **收稿日期**:2010-10-28;**修回日期**:2011-02-20.

* **基金项目**:中国科学院青年人才项目"面向时空轨迹数据的知识发现"(编号:KZCX2-YW-QN303);中国科学院地理科学与资源研究所自主部署创新项目"时空轨迹数据的模式挖掘"(编号:200905004);国家高技术研究发展计划项目"非结构化应急多媒体数据挖掘"(2009AA12Z227)资助.

作者简介:孙嘉(1986-),女,甘肃兰州人,硕士研究生,主要从事空间数据挖掘研究. **E-mail**:sunnie.nju@gmail.com

数据清理的“李德仁法”。空间数据挖掘,是计算机技术、数据库应用技术、空间传感器技术和管理决策支持技术等多学科交叉的边缘学科,其应用正日益渗透到人们认识和改造空间世界的各个学科[6,7]。另一方面,受互联网飞速发展的影响,研究从海量网络数据中获取知识的 Web 挖掘亦成为了数据挖掘领域的新兴学科。Web 挖掘是数据库、信息检索和人工智能(如自然语言处理和机器学习)等学科交叉的新研究领域[8]。一般地,Web 挖掘可以分为3类:Web 超链接结构的挖掘、Web 内容的挖掘和 Web 使用记录的挖掘[9,10]。

近年来,受互联网的逐渐普及和网络中激增的时空数据影响,Web 挖掘和时空数据挖掘共同衍生了一个新的分支——Web 时空数据挖掘。与 Web 挖掘类似,可将 Web 时空数据挖掘分为3类:

基于 Web 超链接结构的时空数据挖掘。传统的 Web 超链接结构挖掘从互联网的组织结构中提取权威页面[9],基于 Web 超链接结构的时空数据挖掘加入空间信息,分析某 Web 服务可以到达的地理范围[11~13]。这方面的研究基于一个假设:如果 A 地区的一个网站引用了目标网站 T,那么说明网站 T 的服务范围可以达到 A 地区。这类分析可以帮助网站分析地域关注差异,查找权威服务。

基于 Web 内容的时空数据挖掘。这方面的研究从 Web 页面的文本内容及其涉及的时空信息中提取知识,即 Web 内容的时空知识发现。Web 内容的时空知识发现,通过分析网络新闻或用户在网络上发表的留言、博客等内容,从中推导空间知识[14~20]。并且,人们对 Web 内容中时空知识的研究,促进了地理信息检索的发展。地理信息检索系统提取 Web 网页中的地理信息,对搜索结果按照地理偏好排序,优先显示用户感兴趣地区的搜索结果[21~25]。

基于 Web 访问记录的时空数据挖掘。人们在网络中的搜索、留言等行为可以看作人类行为和网络资源的交互,对访问记录进行研究可以获得人们在网络中的关注热点和关注度的地理分布等知识。基于 Web 访问记录的空间数据挖掘一般采用提取统计特征和线性回归的方法[16,26]。近几年来,这方面的研究逐渐增多,通过分析搜索引擎使用记录,预测流行音乐、电影、游戏、失业率等各种社会现象[15,27~30]。其中,流感趋势的预测和监测方面的研究最多[16,28,31~33],Google 也于 2009 年发布服务 Google Flu Trend(http://www.google.org/flutrends)用来追踪流感趋势。

Web 时空数据挖掘包括 Web 时空信息提取和 Web 时空知识发现两步。即首先从网页中提取时空信息,并且客观描述其时空信息;然后利用数据挖掘技术,提取出 Web 中的空间知识。下面的章节将从 Web 中的时空信息提取和 Web 时空数据挖掘两方面进行详细介绍。

2 Web 时空信息提取

2.1 Web 中的时空信息

Web 网页中的时空信息提取是 Web 时空数据挖掘的第一步。Web 中的时空信息与普通的时空信息不同,具有以下特点。

(1) *Web 中的时空信息是海量的*。互联网作为目前社会最主要的媒介之一,其中含有大量的时空信息。如时事新闻发生的时间、地点,网站服务器登录日志中的时间、地点,用户发表的博客、微博中记录的时间、地点等。

(2) *Web 中的时空数据是结构性与非结构性的混合*。Web 中包含文本、多媒体、超链接多种文件类型,这些文件中均含有空间信息,在 Web 时空数据挖掘的过程中需要进一步整理。

(3) *Web 中的时空信息具有更新自主性*。Goodchild 认为,网络中的用户可以看作不同的传感器,他们自发地为网络提供、更新各种信息[1]。因此,网络中的空间数据是多人参与的数据集。如 Wikimapia 就是一个描述地球的开放式百科全书,和 Google Earth 相同,用户可以自发地在地图上添加、更新标注。

(4) *Web 中的信息具有不确定性*。Web 中时空信息更新的自主性使得网络中的空间信息激增。海量免费的空间数据虽然蕴含着大量的时空信息,但是其中的不确定性却需要人们进一步甄别和改善[34]。网络中信息的不确定性主要表现在真实性、精度、偏差3方面。首先,用户所提供的信息其真实性难以保证。如2003年3月29日,内地网络媒体发生“比尔盖茨被暗杀”的假新闻事件,引起轰动。近年来,网络中的假新闻越来越多,足可见网络中信息的真实性应当受到关注。其次,精度的不确定性体现在,相对于其他数据,网络数据不能保证足够的精度。由于网络中的信息由大众自发上传和更新,信息来源不同导致了信息的精度不同。再次,网络中的信息可能存在偏差。如在地震发生后,人们的博客记录能够反映地震位置、灾情等信息,但是由于

各人视角不同,记录的信息可能存在偏差。

2.2　Web 时间信息提取

在对网络信息检索的研究中,对时间信息提取的研究非常重要[35]。提取 Web 中的时间信息,对 Web 时空数据挖掘尤为关键。与 Web 中的空间信息相比,时间信息的提取相对简单。Web 中的数据可以划分为 3 类,即超链接数据、网页内容数据和用户使用记录数据。在这 3 类不同的数据中,除了超链接数据外,其他 2 种数据均含有时间信息。网络中的用户使用记录数据多为结构化的数据[36]。因此,用户使用记录中时间信息的提取较为容易。表 1 为文献[36]中示例的用户使用记录数据,其中,时间列准确记录了用户的网络行为发生的日期和时间信息。从网页内容数据中提取时间信息,即文档的时间信息提取(Temporal Information Extraction, TIE)。TIE 做为命名实体识别的一个分支,其研究起始于 20 世纪 90 年代。典型的 TIE 算法一般包括 2 个步骤,时间的表示和识别该时间发生的相关事件[37]。

时间的表示方法由 Allen 提出,用代数的区间来表示时间。这是因为文档内容的时间信息分为相对时间和绝对时间 2 种[38],如“2010 年 7 月 8 日”为绝对时间,“明天”、“后两天”等这一类描述时间的名词为相对时间。识别与时间相关的事件的方法主要有自动注解[39]、概率模型[37]和粗糙集[40]等。其中,文献[40]将事件的名称作为输入,收集与该事件相关的网页,输出事件发生的可能的时间段。

2.3　Web 空间信息提取

网页中的地理信息主要分为 2 类:源位置(source geography)和目标位置(target geography)[41]。源位置包括网页服务器的物理位置、网页提供者的地理位置、用户的 IP 位置等;目标位置指网页内容中与网页主题有关的地理位置。对于某些网页,它的源位置和目标位置相同,如某公司或商店的网站。同时,源位置和目标位置也可以不同。如 CNN 网站上一篇关于北爱尔兰的文章,它的源位置为美国但是目标位置是英国。与目标位置相比,源位置(如 IP 地址、主机地址)更容易确定。WHOIS 数据库是用来查询域名的 IP 以及所有者信息的传输协议[11],主机位置可以通过该数据库确定[42]。目标位置的提取,主要通过 2 步:地理文本分析(geoparsing)和地理信息标识(geocoding)[24]。

表 1　网络使用日志数据示例[36]

Table 1　An example of Web log file[36]

#	IP 地址	用户名	访问时间	访问方法/地址	状态	大小	来源	代理
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	GET A.html HTTP/1.0	200	3 290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	GET B.html HTTP/1.0	200	2 050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	GET L.html HTTP/1.0	200	4 130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	GET F.html HTTP/1.0	200	5 096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	GET A.html HTTP/1.0	200	3 290	-	Mozilla/3.04 (Win95, I)
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	GET B.html HTTP/1.0	200	2 050	A.html	Mozilla/3.04 (Win95, I)
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	GET R.html HTTP/1.0	200	8 140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	GET C.html HTTP/1.0	200	1 820	A.html	Mozilla/3.04 (Win95, I)
9	123.456.78.9	-	[25/Apr/1998:03:10:02 -0500]	GETO.html HTTP/1.0	200	2 270	F.html	Mozilla/3.04 (Win95, I)
10	123.456.78.9	-	[25/Apr/1998:03:10:45 -0500]	GET J.html HTTP/1.0	200	9 430	C.html	Mozilla/3.04 (Win95, I)
11	123.456.78.9	-	[25/Apr/1998:03:12:23 -0500]	GET G.html HTTP/1.0	200	7 220	B.html	Mozilla/3.04 (Win95, I)
12	123.456.78.9	-	[25/Apr/1998:05:05:22 -0500]	GET A.html HTTP/1.0	200	3 290	-	Mozilla/3.04 (Win95, I)
13	123.456.78.9	-	[25/Apr/1998:05:06:03 -0500]	GET D.html HTTP/1.0	200	1 680	A.html	Mozilla/3.04 (Win95, I)

2.3.1　地理文本分析

地理文本分析是指从非结构化或半结构化的文本中提取地理信息的过程。地理文本分析的过程通常分为 3 步:在页面中抓取地名实体、为地理名词去歧义、确定正确的地名或位置[12, 43]。其中,去歧义是地理文本分析的关键。据统计,37% 的地理名词都含有歧义[42]。地理名词的歧义表现在 2 个方面:非地理名词的歧义和地理名词的歧义。前者指地理名词具有非地理意义,如 Turkey 既可以指土耳其,又有火鸡的意思。后者指同一个地理名词可能同时指示 2 个地方,如 London 既可以指英国的首都,也是美国肯塔基州的伦敦[44]。地理文本分析最常用的方法是命名实体识别,除此之外,基于统计和基于本体的识别方法也是地理文本分析的常用方法[34]。大多数命名实体识别算法结合自然语言处理算法和地名辞典识别地名,利用已知地名和人类语言习惯

从上下文中提取地理信息[34,41]。这种方法与其他语义识别的方法相比较为简单,只需要确定地名去歧义的准则。如当地名出现地理歧义时,选择人口较多的地区[34]或当文档中出现多个地名时,选择出现频率最高的地区[45]。利用命名实体识别进行地理文本分析的过程如下[12,41,44,46,47]:

(1) 在 Web 上下文中收集地名辞典中出现的名词。

(2) 利用自然语言处理的算法,去除一部分非地理意义的名词,如名词前面有 Mr. 的为非地理名词,名词前为 city of 的为地理名词。

(3) 对其他有歧义地名的各语义设置初始置信度,当一个地名指向多个地区时,根据各地区的人口、隶属区域计算置信度。

(4) 设置置信度阈值,确定每个地名指向的目标位置。

命名实体识别方法的优点是算法简单,不需要选择训练集,其缺点是这种方法无法识别地名辞典外的地理名词。基于统计方法的地理文本分析采用机器学习的方法,定义训练集,计算代价较高[48~50]。基于本体的识别方法通常应用在地理搜索引擎中,用于识别商业对象或者本地搜索。对于普通的 Web 网页,命名实体识别的方法已经足够。

2.3.2 地理信息标识

地理文本分析得到了地理对象结构化的表达,地理信息标识是对地理对象赋予地理标识元数据的过程。地理信息标识分为 2 步:匹配和定位。即首先将提取出的地理要素(如邮编、区号、城市名),与地名辞典中的地名匹配,然后利用地名辞典中的信息对地理对象进行定位,将结构化的文本信息转化为地理坐标信息[51]。人们通过地理文本分析和地理信息标识得到 Web 中的空间信息,进而用空间的方式管理、存储和发现网络中文本、图片等文件。同时,Web 中丰富的地理信息也为 Web 中的时空数据挖掘提供了大量的素材。

3 Web 时空数据挖掘

获得了 Web 中的空间信息后,如何从 Web 的空间信息中抽取感兴趣的、有价值的隐含信息,成为人们关心的问题。Web 时空数据挖掘就是将传统的空间数据挖掘技术和 Web 中的空间数据结合起来,发现潜在的有用模式和隐藏知识的手段。

3.1 基于 Web 超链接结构的空间数据挖掘

Web 超链接结构的空间数据挖掘从网络的链接关系和 Web 页面的空间信息中推导知识,分析网站的服务范围,寻找权威服务。这类方法通过分析引用目标页面的所有网页的空间分布,加入地名辞典、用户使用日志等信息,得到某 Web 服务可以到达的地理范围[11,12,21]。服务范围计算方法基于一个假设:如果处于 l 地区的一个网站引用了另一个网站 w,那么说明网站 w 的服务范围可以达到 l 地区[11]。那么,该网络资源的服务范围在地图上表示为以地区 l 的几何中心为圆心的一个圆。一般地,l 并不唯一,网站 w 的服务范围用多个半径不同的圆表示,圆的半径为:

$$\mathrm{Radium}(w,l)=\mathrm{Links}(w,l)/\mathrm{Page}(w) \quad (1)$$

其中,$\mathrm{Links}(w,l)$ 为地区 l 中链接到网站 w 的页面数;$\mathrm{Page}(w)$ 为引用目标网站 w 全部的链接数。这种方法比较直接,但是得到的结果不能反映地区 l 的子区域对该网络资源的关注程度。Ding 等[52]对此方法进行了改进,加入树状的地名辞典信息,通过定义 Power 和 Spread 来表示网站 w 的服务范围。Power 用来度量网络资源 w 在地区 l 中的受关注程度,Power 值高的范围为候选服务范围(CGS):

$$\mathrm{Radium}(w,l)=\mathrm{Links}(w,l)/\mathrm{Page}(l) \quad (2)$$

其中,$\mathrm{Links}(w,l)$是地区 l 的网页中链接到网页 w 的页面数,$\mathrm{Page}(w)$是指地区 l 中的所有网页数。

Spread 表示地区 l 的子区域对网站 w 关注的均一程度,有矢量、相对误差、熵 3 种计算方法,但是意义都相同。如果地区 l 中,各个子区域对网站 w 的关注程度一致,那么 Spread 最大。当某地区的 Power 和 Spread 都比较高时,则认为该区域属于网站 w 的服务范围。但是众所周知,网络资源的服务范围和受关注程度并不仅仅由 Web 页面的超链接结构决定,也与点击该 Web 页面用户的地理位置有关。Wang 等[12]提出了另一种网络资源服务范围的计算方法,在文献[52]的基础上,提出类似 PageRank 的算法,结合该网络资源访问用户的地理位置、网页的目标位置,对每个区域和其子区域计算权重,得到 Web 网页的服务范围。

综上所述,基于 Web 超链接结构的空间数据挖掘的各种方法,在分析 Web 网页超链接结构的基础上,不断加入各种信息,如树状结构的地名辞典、用户的访问信息等,使人们对网络资源服务范围的计算更加精确。

3.2 基于 Web 内容的时空数据挖掘

基于 Web 内容的空间数据挖掘从网页的页面内容中提取时空知识,通常指从用户发表的网络博

客和留言中推导空间知识。需要注意的是,基于 Web 内容的时空数据挖掘技术促进了地理信息检索(Geographic Information Retrieval,GIR)技术的应用和发展。

网络中储存着各种 Web 新闻网页、用户博客等。其中,网络博客已成为目前网络中广泛的网页形式之一。这些网络博客的内容与作者年龄、地理位置等个人信息有紧密的联系,同时,对于突发的重大事件,与之相关的用户博客或者留言也可以作为重要的消息来源[14~16]。

从网页内容中提取空间知识,最简单的应用是提取 Web 网页中的目标位置并且在地图上可视化显示。但是人们并不满足于简单的地图显示,近年来,一些研究通过对网络博客、微博中用户发表的信息进行时空数据挖掘、推导空间知识。常用的方法有空间关联规则挖掘和基于概率模型的时空数据挖掘[14~20]。基于 Web 内容的空间关联规则挖掘,按照某一主题,对用户的博客内容进行关联规则挖掘,得到人们在某地区的频繁行为模式[20]。主要步骤为:

(1) 通过关键字搜索某一主题下用户发表的博客。如搜索主题涉及"日本旅游"的所有博客。

(2) 对博客内容进行语义分析,将文章分解为句子,从句子中提取事务(transaction)。一个完整的事务包括时间、地点、名词和动词。

(3) 关联规则挖掘。与 APRIORI 算法相似,基于 Web 内容的时空关联规则挖掘主要提取 3 种模式:[时间、地点]→[动词];[时间、地点、动词]→[名词];[时间、地点]→[动词、名词]。

(4) 结果可视化,将频繁项标注在地图上。

这种方法可以对用户的旅游博客进行分析,得到特定地区的频繁旅游模式,为旅游者提供引导,也可以为当地旅游业发展提供参考。从网页内容中提取空间知识的另一种方法是用概率模型从网络博客和微博记录中提取空间知识,主要应用在流感疫情的监测和地震、飓风等重大突发性灾难事件的网络服务中。Ushahidi(http://www.ushahidi.com/)就是一个基于 Web 内容的知识发现的开源平台,用户可以利用手机、邮件、网站向平台上传危机发生地区的重要消息。Ushahidi 也会自动分析 Twitter、Facebook 等网络消息并以可视化的方法呈现在地图上。以海地地震为例,Ushahidi 在地震发生后迅速建立了一个页面,用户可以看到各种救援的需求报告和灾情的描述。图 1 是 Ushahidi 海地地震的紧急信息服务平台(http://www.haiti.ushahidi.com)。

一般地,基于概率模型从网络博客和微博记录中提取空间知识的步骤如下:

(1) 自动从博客中抽取主题和关键字。

(2) 建立时空概率模型:将每篇文档看作一个三元组,那么文档集 C 为一个三元组序列:$C=\{(d_1,t_1,l_1),(d_2,t_2,l_2),\cdots,(d_n,t_n,l_n)\}$,其中 d_i 为第 i 篇文档,$t=\{t_1,t_2,\cdots,t_n\}$ 代表时间序列,$l=\{l_1,l_2,\cdots,l_n\}$ 代表文档集 C 中涉及的地理位置;文档表示为一系列关键字的集合;在时空概率模型中,文档的主题 θ 用关键字 ω 的概率分布 $p(\omega|\theta)$ 表示,常用的时空概率模型有混合模型、指数模型等[14~16]。

(3) 对时空概率模型估计参数,通常选择 EM 算法。

(4) 结果的可视化。一般通过时间或者空间上时空概率模型的"快照"表示,也可以利用相关分析、滤波等方法,查找关注热点中心随时间的变化,对每个时段各地区用户发表博客的情况进一步分析。

从 Web 中提取时空知识,有助于人们了解全球范围或某地区内,网民或大众对某事件的关注程度、关注侧面和某事件的时空频繁模式。目前,地理信息检索已经成为信息检索领域重要的研究方向[21~25],Google、Yahoo 等搜索引擎公司纷纷推出了自己的地理信息检索系统。地理信息检索帮助人们从因特网中发现与查询关键字语义接近、位置邻近的网页,地理信息检索的应用多针对生活搜索,挖掘邻近地区的商户。但是,网页、博客等也是网络中不可或缺的资源,从中发现空间知识,是 Web 空间数据挖掘中的重要部分。地理信息检索系统的主要步骤是:

(1) 对查询关键字进行地理信息推理,包括对查询关键字去除地理歧义,并且扩展查询词。目前,大多数地理信息检索系统利用地理本体去歧义和扩展查询词。李德仁等[53]认为,地理本体有利于有效整合互联网环境下的空间信息资源。本体的概念最先源于哲学领域,是用于描述或表达某一领域知识的一组概念或术语[54]。地理本体,就是对地理对象提供一个关于术语和结构的模型[21],地理本体中存储了地理对象的位置、拓扑关系等特征。目前,有一些研究考虑对网站定义本体,但是这种本体的表示多针对网站的内容和结构,用于表示网站的个性化特征[54,55],不包含地理信息。在地理信息检索中,

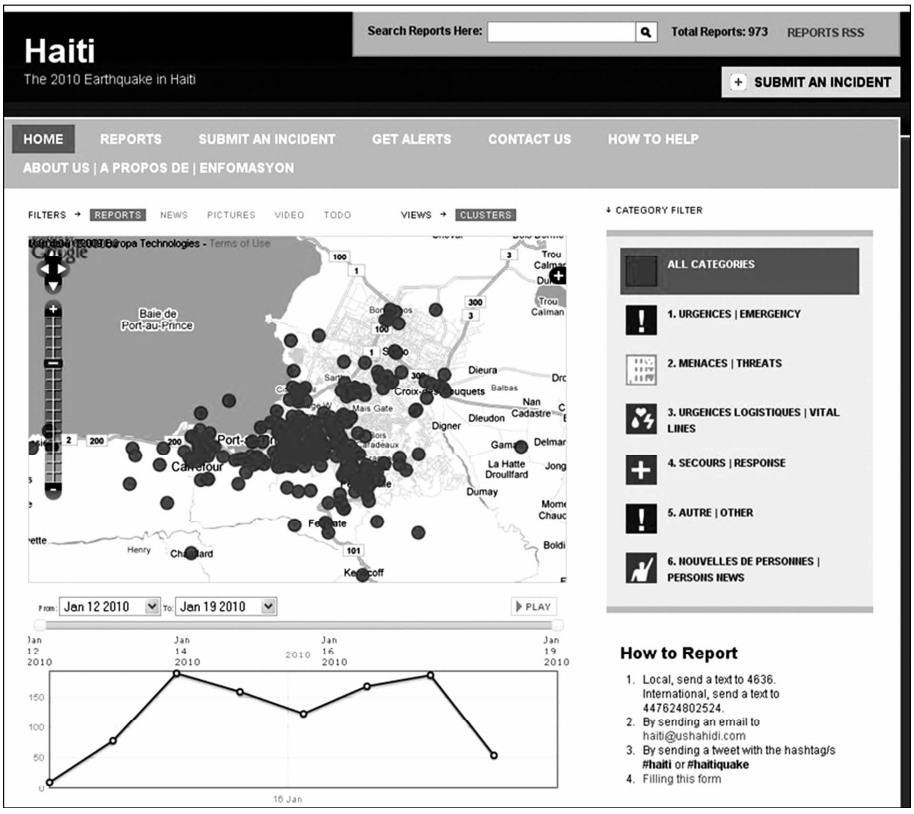

图 1 Ushahidi 海地地震的紧急信息服务平台

Fig. 1 Haiti crisis map in ushahidi. com

主要的地理本体有 Georeference[21]、OnLocus[23,43]、SPIRIT[13]等。

(2) 根据去歧义和扩展后的关键字收集相关网页。

(3) 基于文本相关性和地理相关性对搜索结果排序,显示搜索结果。地理信息检索对结果的排序结合了文本相关性和地理相关性。文本相关性在传统信息检索中有很多研究,如文档反文档(TF-IDF)模型等,而地理信息相关性则包括空间距离、层次化的包含关系等。

3.3 基于 Web 访问记录的时空数据挖掘

传统的 Web 访问记录挖掘,多应用于网站访问顺序分析[56]和电子商务领域的用户分析[10,36],其最常用的方法是关联规则挖掘。如某电子商务网站通过对网站的访问记录进行关联规则挖掘,发现"50% 购买了 product2 的网民居住在西海岸并且年龄在 20 ~ 25 岁之间"[10]。与传统的 Web 访问记录挖掘不同的是,基于 Web 访问记录的时空数据挖掘,通过分析人们在网络上的使用记录,得到搜索行为在地理空间中的分布密度和地域差异,帮助搜索引擎改善搜索质量。其过程分为 3 步:数据收集、数据预处理和数据分析。数据收集主要收集某段时间内的用户访问记录和点击行为。一般地,对于搜索引擎,一个完整的访问记录包括用户 IP、访问时间、搜索条目、用户浏览顺序、浏览器信息和语言等。数据预处理阶段包括文本分析,数据清洗,提取与主题有关的关键字。数据分析阶段对访问记录进行时空数据挖掘。下面,分别就 Web 访问记录的时间分析、空间分析和时空分析 3 个方面进行论述。

3.3.1 Web 访问记录的时间分析

Web 访问记录的时间分析,可以得到人类的网络活动随时间变化的规律,包括流感的预测和监测[16,26,31 ~ 33]、热门歌曲、电影、游戏的预测[27]、地震的预测[20]以及失业率的预测[28 ~ 30]等。这方面的研究方法主要有 2 种:统计特征的提取和回归分析。统计特征的提取方法包括:统计一段时间内的热门搜索词、每个关键词在某时间段内被搜索的次数、绘制搜索量随时间变化的图表等。Google 趋势就是一个对 Web 访问记录进行时间分析的应用(http://www.google.com/trends)。这种方法仅对访问记录本身进行分析,得到搜索行为在时间上的分布密度,但是不能预测和监控现实行为。要研究访问记录与

现实行为之间的关系,主要的方法是回归分析[15,16,26]。

通常,人们选择线性回归的方法来研究搜索词的访问记录和现实统计数据的相关关系,线性回归的方法主要有一元线性回归和多元线性回归[26]。一元线性回归模型如下:

$$f(y) = \beta_0 + \beta_1 g(x) + \varepsilon \tag{3}$$

其中,x 为某个关键词的搜索量比例,y 为目标变量,是现实中的统计数据,如每日疑似病例的数目、每日某电影的票房数等,ε 为误差项。$f(y)$ 和 $g(x)$ 分别为自变量和目标变量的转换函数,可以为 logistic 函数、log 函数等,根据不同的研究目的确定。类似的,多元线性回归的模型如下:

$$f(y) = \beta_0 + \beta_1 g_1(x_1) + \beta_2 g_2(x_2) + \cdots + \beta_n g_n(x_n) + \varepsilon \tag{4}$$

其中,$x = \{x_1, x_2, \cdots, x_n\}$ 可以为 n 个不同关键字的搜索量比例,也可以为某个关键词的搜索量比例和其他与目标变量有关的自变量。Google 在 2009 年推出了流感疫情监测的应用,发现搜索流感相关主题的人数与实际患有流感症状的人数之间存在着密切的关系,并且相对于疾控中心的统计数据,该应用可以提前两周预测到大规模疫情的爆发[26],这为通过网络访问记录预测流感疫情提供了研究方向[27,32,33,58]。

3.3.2　Web 访问记录的空间分析

搜索引擎的访问记录不仅记录了搜索的时间和关键词,还记录了用户的 IP 地址。研究者通过分析不同地区用户的访问记录,得到搜索行为在地理空间的分布密度和地域差异[59]。Web 访问记录的空间分析方法大致分为 2 种:提取统计特征的方法和基于概率的分析方法[59]。提取统计特征的方法对不同区域统计某关键词的搜索量比例,用可视化的方法表现在地图上,直观表示搜索行为的地域差异。Google 搜索解析就是类似的服务,通过分析用户在 Google 中搜索过的条目,将分析的结果显示在世界地图上,表示该搜索条目的关注度差异。提取统计特征将每个区域视为一个整体,表现区域间的关注度差异,但是不能反映搜索行为的热点中心。基于概率的方法弥补了这一不足,既能反映各地区的差异,又能反映用户搜索的热点区[59]。基于概率模型的方法将研究区划分为网格单元,每个网格单元记为 x,对每个网格单元定义概率 p_x。距离热点网格 z 越近,p_x 越大;随着 x 与 z 距离的增大,p_x 逐渐减小,p_x 减小的速度可以用发散程度 P 表示:

$$P = Dd^{-\alpha} \tag{5}$$

其中,C 为中心的频率,α 为 p_x 减小的速度。用迭代的方法求出中心点和发散区域的最大似然值。同时,这一方法发现,热点区与关键词的自然中心一致[59]。

这两种方法的相同之处在于它们都分析了某个关键字下搜索量的空间分布,提取统计特征的方法对于一般的网络应用已经足够,基于概率模型的方法可以得到精细的热点区位置,适用于舆情监控等方面的研究。综上所述,Web 访问记录的空间分析,研究方法由简单的统计数据分析到基于概率模型的空间分析,从发现热点区域到发现热点中心,从发现一个热点中心到发现 K 个热点中心。未来,这方面的研究趋势将是通过 Web 的访问记录,自动发现空间上隐含的所有热点区。

3.3.3　Web 访问记录的时空分析

事实上,针对 Web 访问记录的空间分析和时间分析是密不可分的。由于 Web 访问记录本身带有时间属性和空间属性,Web 访问记录的时间、空间分析都是 Web 访问记录在某地区或某时间段上的“快照”。因此,Web 访问记录的时空分析常用的方法是分时段统计或分地区统计。分时段的统计方法,统计各时段所有地区的用户关注度,分析关注中心随时间的变化情况。Google 搜索解析就结合了时空分析的服务,对每个时间段生成一张搜索行为地理分布的地图,图 2 是 Google 搜索引擎的一个例子,搜索关键词为地震,地理范围为中国,时间为 2008 年。另外,由于 Web 访问记录中蕴含的时空知识与现实中的行为有着密不可分的关联,并且其访问记录数据更新速度快,因此,相对于疑似病例数等精度高但是获得时间较长的统计数据,Web 访问记录数据可在较短的时间内分析出相对真实的结果。

4　总结与展望

Web 时空数据挖掘是 Web 挖掘和空间数据挖掘相结合而发展起来的一个新兴学科。它的出现极大地推动了 Web 挖掘的发展,为我们更加全面、深入地认识互联网、理解人类活动提供了重要的数据和方法基础。目前,通过 Web 时空数据挖掘,人们可以从发现网络资源的地理服务范围、热点话题的区域关注差异和关注热点中心的变化规律等时空知识。作为数据挖掘领域的一个分支,Web 时空数据挖掘应当引入数据挖掘领域已成功应用的软计算方法(如模糊数学、粗糙集、进化计算等),而不仅满足

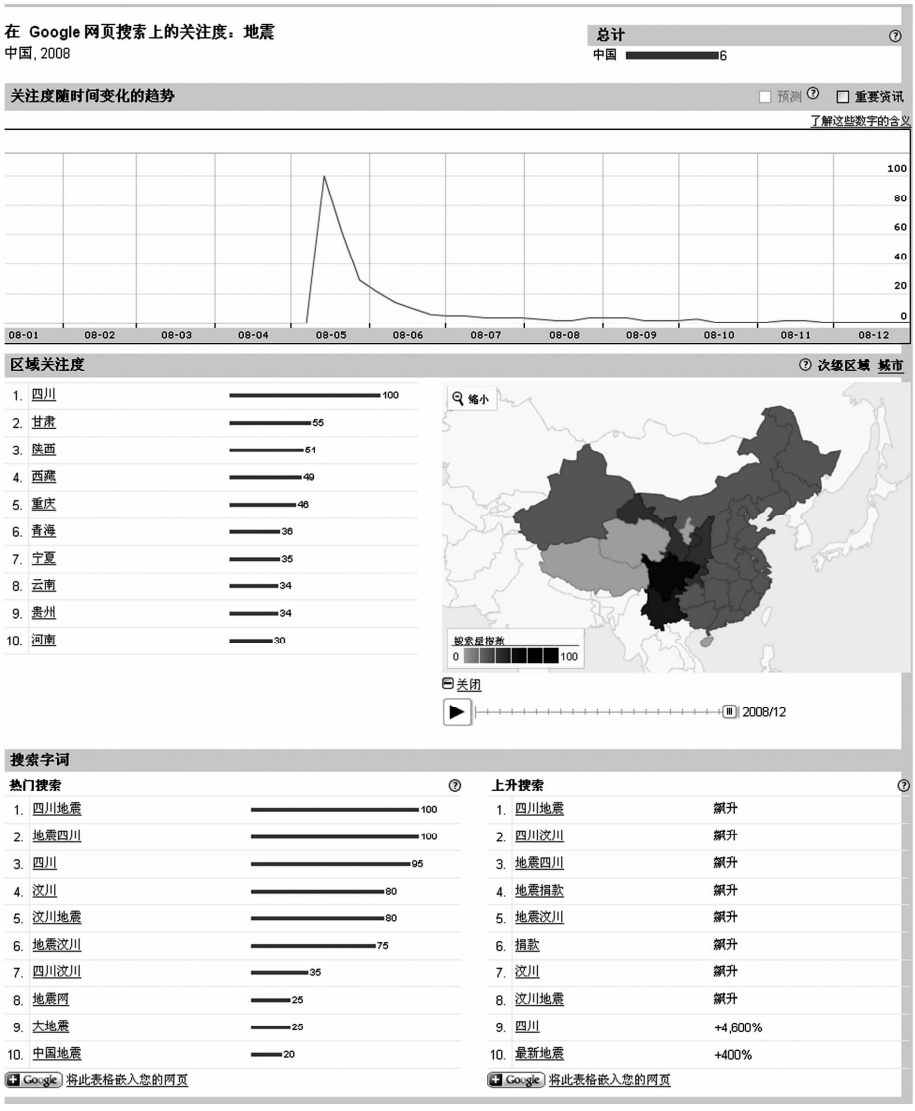

图 2 Google 搜索解析的例子

Fig. 2 The example of Google Insight

于初步的分时段、分地域统计。因此，如何引入在数据挖掘领域已成功应用的一些新算法，对结构复杂的海量 Web 数据进行挖掘，是 Web 时空数据挖掘的方法和应用所面临的主要困难之一。

在互联网、空间数据挖掘、地理移动服务飞速发展的同时，Web 时空数据挖掘的相关应用也将得到新的发展，面临新的机遇和挑战，主要包括 3 个方面：

（1）就空间数据而言，互联网为空间数据的发展提供了新的挑战。Google Earth、WikiMap 等的推出使互联网的用户能够在全球尺度的空间信息服务平台上进行各种研究，享受各种服务，并且这些服务大部分是免费的。互联网中的空间数据不同于传统的空间数据，用户作为认识空间数据的“传感器”，能够参与网络中空间数据的维护和更新。这种互联网地图服务大大地促进了地理信息的应用和普及，并且由此，李德仁等[60]提出了“新地理信息时代”的概念。但是，高度开放的地理信息环境也带来了新的问题，诸如相对困难的地理信息的维护和近年来被许多国家关注的信息安全问题。

（2）随着空间定位信息服务（LBS）的兴起和移动互联网的发展，今后的 Web 时空数据挖掘将不局限于互联网。LBS 服务的关键思想是移动终端发送其位置信息到服务中心，服务中心再通过其数据库找到终端位置附近最相关的信息，然后把信息送回到移动终端[61]。如美国著名的 LBS ——Foursquare，在其出现之时，就以狂飙突进的风头迅速成为了互联网和移动服务的热点。而在移动通信领

域,手机用户越来越多的使用互联网业务和 LBS 服务,运营商、IT 业和互联网巨头将出现越来越多的跨界竞争和合作,这为 Web 时空数据挖掘的发展提供了广阔的空间。未来,Web 时空数据挖掘能够帮助 LBS 探索有效的商业模式,并且通过 LBS 和 Web 空间数据挖掘的有机结合,为政府部门提供决策支持。

(3) 网络和移动通信技术在刻画人与人行为的同时,物与物之间行为的刻画也慢慢发展起来,这就是物联网。物联网通过传感器、全球定位系统(GPS)和网络,将任何一种物品和互联网连接起来,实现物品的智能化识别、定位、跟踪和管理[62],我国也在无锡等城市开展物联网的试点工作。物联网中的数据与互联网中的数据类似,不仅数据量大,而且数据类型复杂多样、数据更新周期较短。而 Web 时空数据挖掘所处理的,正是数据海量、时空特征明显、数据类型多样、更新速度快的 Web 数据,通过对物联网海量数据进行实时在线挖掘,Web 时空数据挖掘能够为物联网用户提供更快速、便捷、油耗的数据查询分析服务,从而评价指导区域经济发展模式,产生新的应用和商业模式。

参考文献(References):

[1] Goodchild M F. Citizens as sensors: The world of volunteered geography[J]. *GeoJournal*, 2007, 69(4):211-221.

[2] Klösgen W, Zytkow J. Handbook of Data Mining and Knowledge Discovery[M]. Oxford: Oxford University Press,2002.

[3] Ester M, Kriegel H P, Sander J. Spatial data mining: A database approach[C]//Proc. of the Fifth Int. Symposium on Large Spatial Databases (SSD 97). Berlin, Germany,1997:47-66.

[4] Han J W, Koperski K, Stefanovic N. GeoMiner: A system prototype for spatial data mining[C]//Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data. New York, NY, USA: ACM, 1997.

[5] Li Deren, Wang Shuliang, Li Deyi. Spatial Data Mining Theories and Applications[M]. Beijing: Science Press, 2006.[李德仁, 王树良, 李德毅. 空间数据挖掘理论与应用[M]. 北京:科学出版社,2006.]

[6] Li Deren, Wang Shuliang, Shi Wenzhong, *et al*. On spatial data mining and knowledge discovery [J]. *Geomatics and Information Science of Wuhan University*, 2001,26(6):491-499.[李德仁, 王树良, 史文中,等. 论空间数据挖掘和知识发现[J]. 武汉大学学报:信息科学版, 2001,26(6):491-499.]

[7] Li Deren, Wang Shuliang, Li Deyi, *et al*. Theories and technologies of spatial data mining and knowledge discovery[J]. *Geomatics and Information Science of Wuhan University*,2002,27(3):221-233.[李德仁,王树良,李德毅,等. 论空间数据挖掘和知识发现的理论与方法[J]. 武汉大学学报:信息科学版, 2002,27(3):221-233.]

[8] Kosala R, Blockeel H. Web mining research:A survey[J]. *ACM SIGKDD Explorations Newsletter*, 2001, 2(1):1-15.

[9] Han Jiawei, Meng Xiaofeng, Wang Jing, *et al*. Research on Web mining: A survey[J]. *Journal of Computer Research and Development*, 2001, 38(4):405-414.[韩家炜, 孟小峰, 王静,等. Web 挖掘研究[J]. 计算机研究与发展, 2001 ,38(4):405-414.]

[10] Cooley R, Mobasher B, Srivastava J. Web mining: Information and pattern discovery on the World Wide Web[C]//9th International Conference on Tools with Artificial Intelligence (ICTAI '97), 1997:558-567.

[11] Buyukkokten O, Cho J, Garcia-molina H, *et al*. Exploiting geographical location information of web pages[C]//Proceeding of the ACM SIGMOD Workshop on the Web and Databases 1999 (Web DB'99), Philadelphia, Pennsylvania: [s. n], 1999:1-18.

[12] Wang C, Xie X, Wang L, *et al*. Detecting geographic locations from web resources[C]//Proceeding of the 2005 Workshop on Geographic Information Retrieval. New York, NY, USA: ACM, 2005:17-24.

[13] Silva M J, Martins B, Chaves M, *et al*. Adding geographic scopes to web resources[J]. *Computers, Environment and Urban Systems*, 2005, 30(4): 378-399.

[14] Mei Q, Liu C, Su H, *et al*. A probabilistic approach to spatiotemporal theme pattern mining on Weblogs[C]//Proceeding of the 15th International Conference on World Wide Web. New York, NY, USA: ACM, 2006:533-542.

[15] Sakaki T, Okazaki M, Matsno Y. Earthquake shakes Twitter users: Real-time event detection by social sensors[C]//Proceeding of the 19th International Conference on World Wide Web. New York, NY, USA: ACM, 2010:851-860.

[16] Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages[C]//Proceeding of the 1st Workshop on Social Media Analytics (SOMA'10). New York, NY, USA: ACM, 2010:32-45.

[17] Mehler A, Bao Y, Li X, *et al*. Spatial analysis of new sources [J]. *IEEE Transactions on Visualization and Computer Graphics*, 2006, 12(5): 765-772.

[18] Xu K, Li R, Bao S H, *et al*. SEM: Mining spatial events from the Web[C]//Proceeding of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer-Verlag, 2008:393-404.

[19] Kurashima T, Tezuka T, Tanaka K. Mining and visualizing local experiences from Blog entries[C]//Proceeding of the 17th International Conference on Database and Expert Systems Applications. Berlin, Heidelberg: Springer-Verlag, 2006, 4 080:213-222.

[20] Kurashima T, Tezuka T, Tanaka K. Blog map of experiences: Extracting and geographically mapping visitor experiences from Urban Blogs[C]//International Conference on Web Information Systems Engineering. Berlin, Heidelberg: Springer-Verlag,

2005, 3 806:496-503.

[21] Christopher B J, Alia I A, David F, *et al*. The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing[J]. *Geographic Information Science*, 2004,3 234:125-139.

[22] Jones C B, Alani H, Tudhope D. Geographical information retrieval with ontologies of place[C]// Proceeding of the International Conference on Spatial Information Theory: Foundations of Geographic Information Science. Morro Bay, CA, USA: Springer, 2001, 322-335.

[23] Borges K A V. Use of an Ontology of Urban Places for Recognition and Extraction of Geospatial Evidences on the Web[D]. Federal University of Minas Gerais,2006.

[24] Larson R R. Geographic information retrieval and spatial browsing[C]// Geographic Information Systems and Libraries: Patrons, Maps, and Spatial Information,1996:81-124.

[25] Makowetz A, Chen Y, Suel T, *et al*. Design and Implementation of a Geographic Search Engine, TR-CIS-2005-03[R]. New York: Polytechnic University, Brooklyn, 2005.

[26] Ginsberg J, Mohebbi M H, Patel R S, *et al*. Detecting influenza epidemics using search engine query data[J]. *Nature*, 2009, 457: 1 012-1 015.

[27] Goel S, Hofman J M, Lahaie S, *et al*. What can search predict [C]// www. 2010, April 26 – 30, 2010, Raleigh, North Carolira, 2010.

[28] Ettredge M, Gerdes J, Karuga G. Using Web-based search data to predict macroeconomic statistics[J]. *Communications of the ACM*, 2005, 48(11): 87-92.

[29] D′amuri F, Marcucci J. "Google it!" Forecasting the US Unemployment Rate with a Google Job Search Index[R]. Bank of Italy, 2009.

[30] Askitas N, Zimmermann K F. Google econometrics and unemployment forecasting[J]. *Applied Economics Quarterly*, 2009, 55(2):107-120.

[31] Quincey E D, Kostkova P. Early warning and outbreak detection using social networking websites: The potential of Twitter[J]. *Electronic Healthcare*, 2010, 27(2): 21-24.

[32] Polgreen P M, Chen Y, Pennock D M, *et al*. Using internet searches for influenza surveillance[J]. *Clinical Infectious Disease*, 2008, 47(11): 1 443-1 448.

[33] Eysenbach G. Infodemiology: Tracking flu-related searches on the web for syndromic surveillance[C]// AMIA Annual Symposium Proceedings,2006:244-248.

[34] Pasley R, Clough P, Purves R S, *et al*. Mapping geographic coverage of the web[C]// Proceeding of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York, NY, USA: ACM, 2008:1-20.

[35] Alonso O, Gertz M, Baeza-Yates R. On the value of temporal information in information retrieval[C]// ACM SIGIR Forum. New York, NY, USA: ACM, 2007,41(2):35-41.

[36] Srivastava J, Cooley R, Deshpande M, *et al*. Web usage mining: Discovery and applications of usage patterns from web data[J]. *ACM SIGKDD Explorations Newsletter*, 2000, 1(2):12-23.

[37] Ling X, Weld D S. Temporal information extraction[C]// Proceedings of the Twenty Fifth National Conference, 2010:1-6.

[38] Allen J. Maintaining knowledge about temporal intervals [J]. *Communications of the ACM*, 1983, 26(11):832-843.

[39] Verhagen M, Gaizauskas R, Schilder F, *et al*. Semeval-2007 task 15: Tempeval temporal relation identification[C]// 4th International Workshop on Semantic Evaluations, Stroudsburg. PA, USA, 2007:75-80.

[40] Schockaert S. Reasoning about Fuzzy Temporal and Spatial Information from the Web [M]. Ghent: Ghent University, 2008.

[41] Amitay E, Har′el N, Sivan R, *et al*. Web-a-where: Geotagging web content [C]// Proceeding of the 27th annual international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR′04). New York, NY, USA: ACM, 2004: 273-280.

[42] Mccurley K S. Geospatial mapping and navigation of the web[C]// Proceeding of the 10th International World Wide Web Conference (WWW′10). New York, NY, USA: ACM, 2005: 221-229.

[43] Borges K A V, Laender A H F, Medeiros C B, *et al*. Discovering geographic locations in Web pages using urban addresses [C]// Proceeding of the 4th ACM Workshop on Geographical Information retrieval. New York, NY, USA: ACM, 2007:31-36.

[44] Li H, Srihari R K, Niu C, *et al*. Location normalization for information extraction[C]// Proceeding of the 19th CoLING. Morristown, NJ, USA: Association for Computational Linguistics, 2002:1-7.

[45] Tezuka T, Tanaka K. Landmark extraction: A Web mining approach [J]. *Spatial Information Theory*,2005,3 639(2 005): 379-396.

[46] Rauch E, Bukatin M, Baker K. A confidence-based framework for disambiguating geographic terms [C] // Proceeding of the HLT-NAACL 2003 Workshop on Analysis of Geographic References. Morristown, NJ, USA: Association for Computational Linguistics, 2003:50-54.

[47] Li H, Srihari R K, Niu C, *et al*. InfoXtract location normalization:A hybrid approach to geographic references in information extraction[C]// Proceeding of the HLT- NAACL 2003 Workshop on the Analysis of Geographic References. Morristown, NJ, USA: Association for Computational Linguistics, 2003, 1:39-44.

[48] Burger J D, Henderson J C, Morgan W T. Statistical named entity recognizer adaptation[C]// Proceedings of CoNLL-2002. Morristown, NJ, USA: Association for Computational Linguistics, 2002:163-166.

[49] Malouf R. Markov models for language-independent named entity recognition[C]// Proceeding of CoNLL-2002. Morristown, NJ, USA: Association for Computational Linguistics, 2002:187-190.

[50] McNamee P, Mayfield J. Entity extraction without language-specific resources [C] // Proceeding of CoNLL-2002. Morristown, NJ, USA: Association for Computational Linguistics, 2002:183-186.

[51] Delboni M T, Borges K A V, Laender A H F, *et al*. Semantic expansion of geographic web queries based on natural language positioning expressions [J]. *Transactions in GIS*,2007, 11(3): 377-397.

[52] Ding J, Gravano L, Shivakumar N. Computing geographical scopes of web resources[C]//Proceeding of the 26th International Conference on Very Large Data Bases (VLDB'00). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 2000: 545-556.

[53] Li Deren, Cui Wei. Geographic ontology and SIMG [J]. *Acta Geodaetica et Cartographica Sinaica*, 2006, 35(2):144-148. [李德仁, 崔巍. 地理本体与空间信息多级网格[J]. 测绘学报, 2006 , 35(2): 144-148.]

[54] William S, Austin T. Ontologies [J]. *IEEE Intelligent System*, 1999, 1(2):18-19.

[55] Wong T L, Lam W. Learning to refine ontology for a new Web site using a bayesian approach[C]//Proceedings of the Fifth SIAM International Conference on Data Mining. 2007, 7(1):1-12.

[56] Eirinaki M, Vazirgiannis M, Varlamis I. SEWeP: Using site semantics and a taxonomy to enhance the Web personalization process[C]//Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003:1-10.

[57] Pei J, Han J W, Mortazavi-Asl B. Mining access patterns efficiently from web logs [J]. *Knowledge Discovery and Data Mining*,2000, 2000(1 805):396-407.

[58] Brownstein J S, Freifeld C C , Madoff L C. digital Disease detection-harnessing the Web for public health surveillance[J]. *The New England Journal of Medicine*, 2009, 360: 3 153-3 159.

[59] Backstrom L, Kleinberg J, Kumar R, *et al*. Spatial variation in search engine queries[C]//Proceeding of the 17th International Conference on World Wide Web. New York, NY, USA: ACM, 2008: 357-366.

[60] Li Deren, Shao Zhenfeng. A new era of geographic information [J]. *Science in China*(*Series F*), 2009, 39(6):579-587. [李德仁, 邵振峰. 论新地理信息时代[J]. 中国科学:F 辑, 2009, 39(6):579-587.]

[61] Chen Feixiang, Yang Chongjun, Shen Shengli, *et al*. Research on mobile GIS based on LBS[J]. *Computer Engineering and Applications*, 2006, 42(2):200-210. [陈飞翔, 杨崇俊,申胜利, 等. 基于 LBS 的移动 GIS 研究[J]. 计算机工程与应用, 2006, 42(2):200-210.]

[62] Meloan S. Toward a global "Internet of Things"[J]. *Sun Developer Network*,2003:203-227.

Review of Research Progress in Web Spatio-temporal Data Mining

Sun Jia[1,2], Pei Tao[3], Gong Xi[2,3], Zhou Chenghu[1,3]

(1. *Yantai Institute of Coastal Zone Research*, *Chinese Academy of Sciences*, *Yantai* 264003, *China*;
2. *Graduate University of Chinese Academy of Sciences*, *Beijing* 100049, *China*;
3. *Institute of Geographic Sciences and Natural Resources Research*, *Chinese Academy of Sciences*, *Beijing* 100101, *China*)

Abstract: Web Spatio-temporal information describs the spatio-temporal scope of events or actors. One could find spatio-temporal knowledge such as the services scope of network resources, geographical distribution of search behavior, and the web-page-based disaster description. This paper systematically reviews the web spatio-temporal data mining technology and services. Firstly, this paper introduces the unique characteristics of web spatio-temporal data, discusses the methods of web spatio-temporal information extraction. Then, it introduces each type of web spatio-temporal data mining methods. Finally, some challenges and future directions are discussed.

Key words: Web spatio-temporal data mining; Spatial data mining; Spatio-temporal information; Geographic information extraction.